1 Project title

Alex Lim, Edward Ho, Roy Chao, Robin Lee

1.0.0.1 Author contributions

Roy Chao developed and answered question 1.

Robin Lee developed and answered question 1.

Alex Lim developed and answered question 2.

Edward Ho developed and answered question 2.

1.0.0.2 Abstract

As devout fans of professional sports, especially tennis, we discovered and analyzed a dataset that recorded every single ATP (Association of Tennis Professionals) match throughout the past decade. By avidly following these tournaments each year as enthusiasts ourselves, we quickly perceived various factors--such as home advantage, height, and weight--that were extremely prevalent in tennis. Because our dataset featured match and player characteristics, we wondered: what actually played into the outcomes of these matches? Furthermore, we wondered how bettors made their decision with these seemingly unpredictable matches, raising our second inquiry: the accuracy of betting odds. Following our analysis, we noticed that betting odds appeared to be a relatively accurate predictor for winning, as the probability of winning based off lower betting odds was almost identical to the theoretical probability of winning. For our other question, we noticed that only a few of the predictors proved to be statistically significant in predicting win percentage.

1.1 Introduction

1.1.1 Background

The ATP, or Association of Tennis Professionals, is a world-wide male tennis organization that holds various tournaments in a bracket-style structure, which are broken down into various categories: the ATP Finals, ATP Masters 1000, ATP 500, ATP 250, and Grand Slam. With three different skill tiers, these matches feature a diverse collection of professional players, each representing their respective nations. Within these highly intense and competitve tournaments, the athletes compete for a prize pool that holds varying degrees of international significance and prestige, with rewards ranging from monetary earnings to physical accolades (trophies, medals, etc).

In the ATP matches, odds are set by bookmakers, and they are expressed in multiple ways—such as moneyline, fraction, or decimal. In our data, odds are expressed as a decimal. For example, if a player has odds of 1.6 and we bet 1 dollar on him to win, we will get 1.6 dollar if he actually wins. Keep in mind that this is the pre-match betting odds, which is why losers typically have higher odds

(as shown in the plot below). If the bookmaker predicts a player has a very low chance of winning a match, they will set high odds for this player. Thus, if we bet on someone with a low chance to win and he does end up winning, we will get a much higher payout.

Our data is composed of the representing players' countries in their respective matches, and the associating average betting odds for the two players in each individual match. In other words, each observations tracks the statistics of two competing players in a single match; this data consists of the country of which they are representing, the odds which they were projected to win the match, and the number of games won in the match. With so much information regarding these ATP matches, we utilized this opportunity to analyze the statistical aspects behind professional sports.

Our primary motivation behind this project is to study the diversity and mystery of sports; in our case, tennis. Ultimately, the outcome of a game is determined by various underlying factors, and we can not associate a win or loss to only one specific reason. Therefore, being able to draw potential associations can be important in predicting and explaining certain outcomes. Since tennis is such a global sport, understanding if a player's ethnic background has influence on their odds of winning and performance during a match is especially valuable. What we hope to learn in our analysis is whether or not we can draw an association between the players' performance with external factors, such as the match location, and if these odds winning are accurate. Answering these questions allows us to know what are some influencing factors on players and how they might be connected.

1.1.2 Aims

For our first question, we wanted to tackle the question of how well betting odds predicted the outcomes of a match. As is with a majority of professional sports--especially tennis--the ATP featured a substantial amount of gambling, with bettors placing varying sums of money before each match based primarily on the odds created by bookmakers prior to each match. With millions of dollars on the line, this raised the question: how much should we believe these predictions? Assuming no bettor deliberately placed their money expecting to lose, we sought to discern how effectively the bookmakers anticipated and accurately predicted the outcomes, and how much bettors were willing to risk on a player with higher bettings odds and a lower probability of winning a match. In order to approach this question, we decided to utilize several graphical representations, comparing the percentage of games won per betting odd with the theoretical probability of winning.

As for our second question, we were primarily interested in exploring is what variables might've had an influence on player performance as well as fitting a regression model and seeing how well it performs. We see that home-court advantage exists in many sports, such as basketball, baseball, football, etc. But does it exist in tennis? Since tennis tournaments happen all around the world and player ethnicity is diverse, we will define home court advantage as a player playing in his home country. It's also a myth that higher tennis players have an advantage and perform better. Is this really the case? And what variables might potentially affect performance? So our approach to this question was designing a multiple linear regression model fitted with various predictor variables, such as odds, height, etc., while our response variable was the percent of games won in a match which is a sufficient measure of player performance. One exploratory method included finding the p-values and comparing them at a 0.05 significant level, we were able to see which

predictors are significant to prognosticating each individual player performance. It's important to note that there are games within a match, so one match requires at least 12 "games" won before declaring a match winner.

1.2 Materials and methods

The goal of this section is to describe your dataset(s) and sketch out your analysis.

1.2.1 Datasets

For this study, the observational units are ATP matches and the variables are location, WFlag/LFlag, Winner_betting_odds/Loser_betting_odds, W_home/L_home, W_percentgames/L_percentgames, and OddsBracket_W/OddsBracket_L. The matches recorded in the dataset are indicated with the variable descriptions in Table 1.

As with most data sets, there are some missing values that we cannot account for. In our analysis, we discovered that 'W_flag' and 'L_flag' contain the most missing values. This is most likely due to the number of players who are less popular and as such, have made fewer appearances in ATP matches. Because of this, little information would be known about these players, specifically their country of origin.

> **Table 1:** descriptions, units, and number of missing values for each variable in the dataset.

Variable Name	Description	Type	Missing Data
Location	Country where match is played	Categorical	0
WFlag	Country of winner	Categorical	40
LFlag	Country of loser	Categorical	566
Winner_betting_odds	Odds of winning bet (pre-match)	Numeric	0
Loser_betting_odds	Odds of losing bet (pre-match)	Numeric	0
W_home	Indicates whether the winner played the match in his home country (1=yes, 0=no)	Indicator	0
L_home	Indicates whether the loser played the match in his home country (1=yes, 0=no)	Indicator	0
W_percentgames	Percentage of games won in the match for winner	Numeric	1
L_percentgames	Percentage of games won in the match for loser	Numeric	1
OddsBracket_W	Binned Winner Betting Odds	Intervals	0
OddsBracket_L	Binned Winner Betting Odds	Intervals	0

	Location	W_flag	L_flag	W_betting_odds	L_betting_odds	W_home	L_home	W_percentgames	L_percentgames
0	AUS	KOR	JPN	1.655	2.280	0	0	0.800000	0.200000
1	AUS	BRA	GER	2.395	1.605	0	0	0.533333	0.466667
2	AUS	SRB	ESP	1.735	2.150	0	0	0.560000	0.440000
3	AUS	USA	AUS	2.105	1.770	0	1	0.500000	0.500000
4	AUS	FRA	DEN	2.705	1.485	0	0	0.565217	0.434783

For table 2, these are the associated variables which we used to perform our multiple linear regression model. We made sure to exclude any missing data variables so therefore we have no missing values.

Table 2: descriptions, units, and number of missing values for each variable in the dataset.

Variable Name	Description	Type	Missing Data
Odds	These are the associated predicted betting Odds of each player winning their match	Numerical	0
Percent games won	This was the percent of games won per match for each player	Numerical	0
Home	Whether the player is playing in their home country during said match (1 = yes, 0=no)	Indicator	0
Height	Height of players (in CM)	Numeric	0
Rank_Points	Indicates the points each individual player has accumulated from their matches	Numeric	0
Odds_x_Ranking_Points	This is the interaction between Odds and the Ranking Points predictor variables	Numeric	0

	Odds	Percent games won	Home	Height	Rank_Points	Odds_x_Ranking_Points
0	1.655	0.800000	0	180.0	1115.0	1845.325
1	2.395	0.533333	0	183.0	805.0	1927.975
2	1.735	0.560000	0	185.0	1131.0	1962.285
3	2.105	0.500000	0	188.0	812.0	1709.260
4	2.705	0.565217	0	175.0	797.0	2155.885

1.2.2 Methods

Our project is an exploratory analysis to identify which variables are potentially significant in determining the percentage of games won in tennis matches. In order to approach the first question, we divided the odds into multiple bins named OddsBracket_W/OddsBracket_L (with values 1-1.3, 1.3-1.6, etc) which indicates the theoretical likelihood of winning the match and the corresponding percentage of players who won when their odds were in the associated bin. With

our data properly divided in this way, we completed a kernel density estimate of the percentage of matches won against betting odds. Afterward, we designed graphics that displayed how the betting odds of winners favored the betting odds of losers. For the second question, we created a linear regression model, determining how betting odds, home, height, player ranking points, and the interaction term between betting odds and player ranking points (since odds and player ranking points may be dependent on one another) affect a player's win percentage. With this, we constructed a table displaying the R^2 values, coefficients, standards errors, and p-values. Finally, we included a plot of the predicted winning percentage vs the actual winning percentage.

1.3 Results

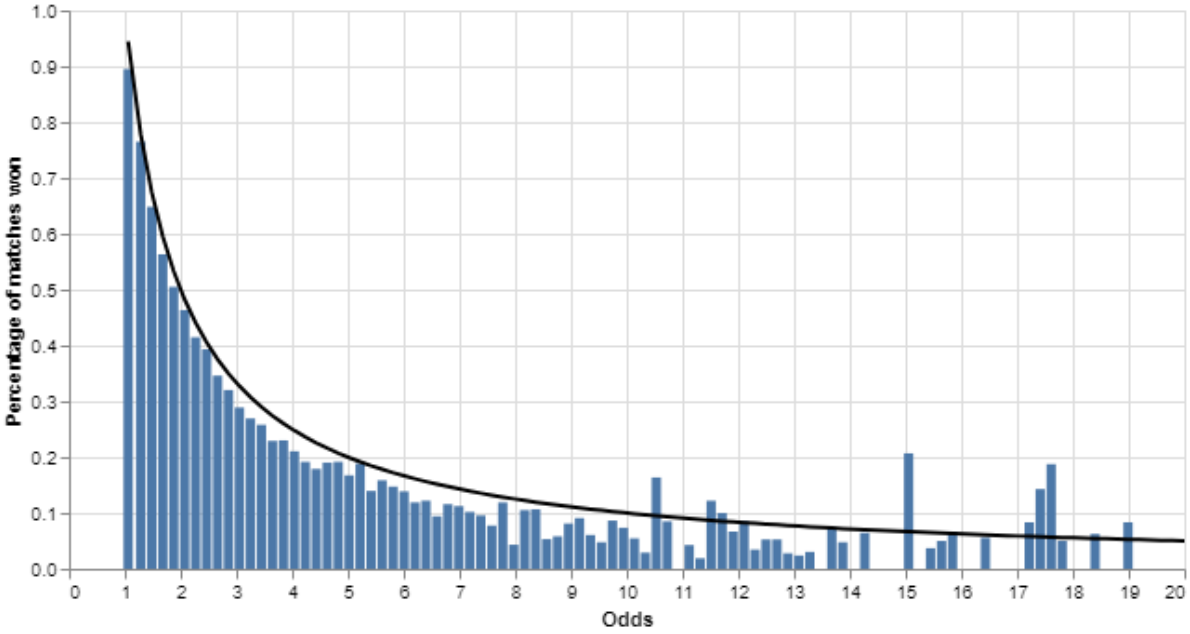
Graphic/Explanation 1

As it was discussed in our first question, we want to analyze the validity of Odds as a valuable predictor of percent of games won. Showing how the bins are depicted below:

Odds	Percent matches won	Theoretical probability of winning
1-1.2	0.888285	0.833-0.999
1.2-1.5	0.727470	0.666-0.833
1.5-1.8	0.575157	0.555-0.666
1.8-2.1	0.486809	0.476-0.555
2.1-2.5	0.416678	0.400-0.476
2.5-3.0	0.338044	0.333-0.400
3.0-3.8	0.261660	0.263-0.333
3.8-5.0	0.198922	0.200-0.263
5.0-7.0	0.140310	0.143-0.200
7.0-10.0	0.084107	0.100-0.143
10.0-15.0	0.051434	0.066-0.100
Over 15.0	0.035112	less than 0.066

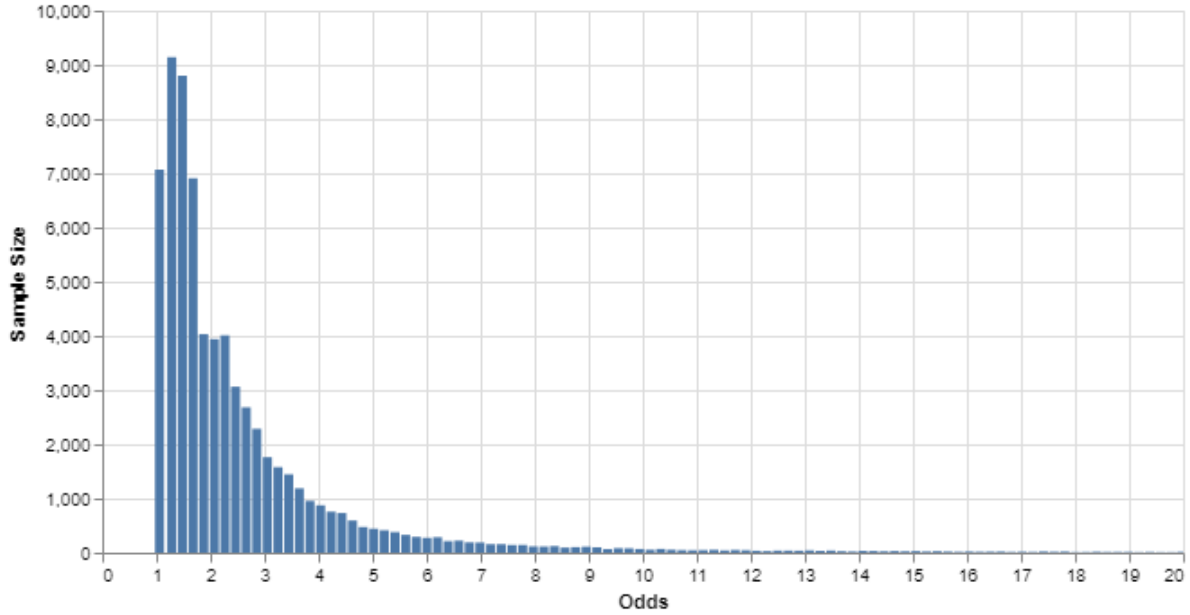
Notice that that most of the values for "Percent matches won" are within the range of the theoretical probability. Some of the values for higher odds are out of the range. This may be due to low sample sizes among the higher odds skewing the data.

Graphic/Explanation 2



This may be easier to visualize using a graph. The graph above displays the percentage of matches won vs. betting odds. Note: the black curve represents the theoretical probability of winning given the odds.

Graphic/Explanation 3

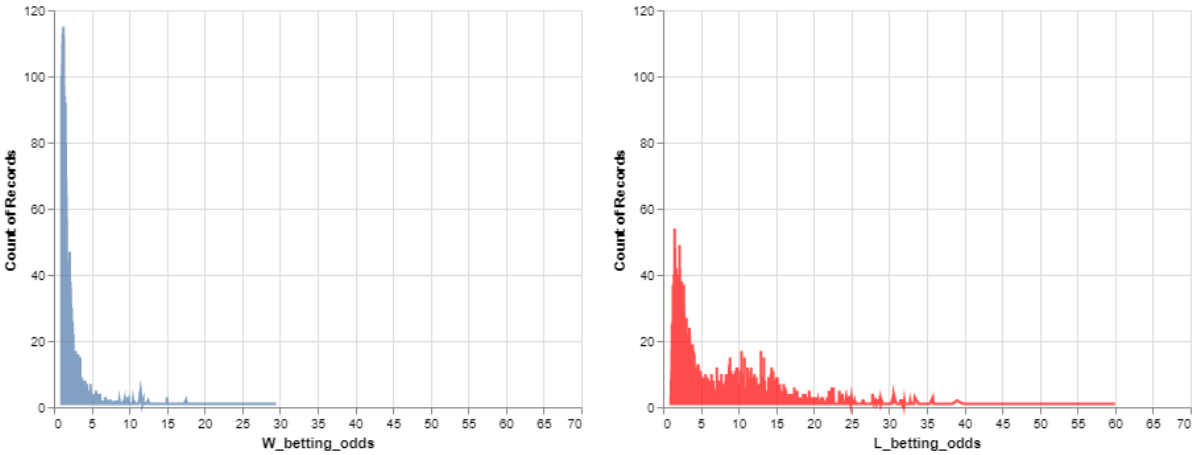


Betting odds seem to be a very good predictor of the probability of winning. For lower odds (1.0-2.5), the percentage of matches won is almost identical to the theoretical probability of winning. For higher odds, betting odds seem to be a less accurate predictor and seem to overestimate the probability of winning. We believe this may be due to low sample sizes for higher odds skewing the data (see graph above). However, the percentage of matches won still follows a decreasing trend as odds increase. Notice the majority of odds in the data are between 1-2, and very few odds are higher than 5.

Graphic/Explanation 4

The model below contrasts the winners' and losers' betting odds, where we see that the plot for

winners' betting odds (the odds of winning a bet pre-match) is a lot more right-skewed than losers' betting odds (the odds of losing a bet pre-match). As we mentioned above, losers typically have higher odds.



Notice that intuitively, from the two graphics, it's already showing correlationship in which losers appear to have higher odds overall than those of winners but lower percentages overall of matches won at respectively, while winners show the inverse.

Graphic/Explanation 5

Notice that there seems to be a trend between the overall matches won and the appearant betting odds, but does this perhaps hold up with other predictor variables when associated with individual games within a match? Below we ran a multiple linear regression model which is fitted with the predictor variables Odds, Home, Height, Rank Points, and the interaction between Odds and Ranking Points as well to predict the response variable percent of games within a match won. Below is our model:

OLS Regression Results

Dep. Variable:	Percent games won	R-squared:	0.167
Model:	OLS	Adj. R-squared:	0.167
Method:	Least Squares	F-statistic:	2744.
Date:	Sat, 19 Mar 2022	Prob (F-statistic):	0.00
Time:	04:36:00	Log-Likelihood:	43714.
No. Observations:	68229	AIC:	-8.742e+04
Df Residuals:	68223	BIC:	-8.736e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5185	0.006	93.187	0.000	0.508	0.529
Odds	-0.0143	0.000	-57.554	0.000	-0.015	-0.014
Home	0.0016	0.001	1.111	0.266	-0.001	0.004
Height	3.259e-05	2.94e-05	1.108	0.268	-2.51e-05	9.03e-05
Rank_Points	1.718e-05	3.9e-07	44.001	0.000	1.64e-05	1.79e-05
Odds_x_Ranking_Points	-3.699e-06	2.15e-07	-17.173	0.000	-4.12e-06	-3.28e-06

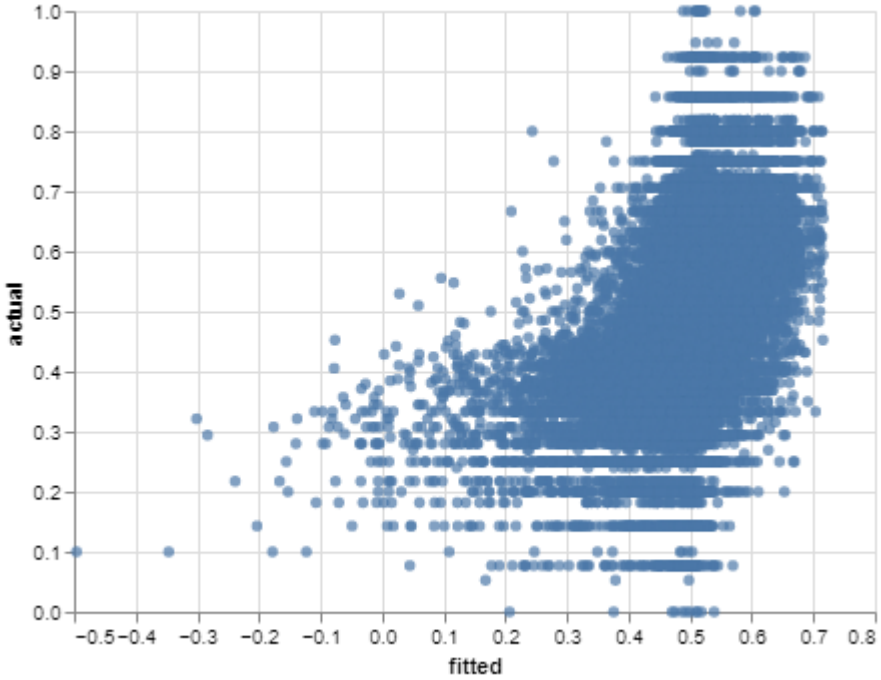
Omnibus:	127.312	Durbin-Watson:	0.862
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.661
Skew:	0.039	Prob(JB):	1.17e-33
Kurtosis:	3.217	Cond. No.	6.19e+04

Initially, you may notice that the coefficient estimates are quite small, but this is because we are predicting a percentage therfore these values are contained between 0 to 1. For example, we can predict for each one-unit increase in Odds, there is about a -0.0143 or 1.43% decrease in percentage of games won within a match.

Additionally, since we want to determine the strength of our model, we also found R^2 which helps determine the proportion of the variability in Y that can be explained using X. Our R^2 produces a value of 0.1674, which is quite low meaning our model didn't fit the data too well and there was quite a bit of noise produced.

With a null hypothesis that our coefficient estimates are not valuable predictors of percentage of games won and at a threshold of 0.05, we find that Odds, Rank Points, and Odds_x_Ranking_Points is statistically significant and thus a good predictors, while Height and Home are poor predictors.

Graphic/Explanation 6
To support this low R^2, we produced a fitted vs. actual graphic below:



Notice that our actual values(y-axis) only located between 0 to 1, with a high density of values from 0.3 to 0.7. But when looking at our fitted(x-axis), most of the data is contained between 0.1 and 0.7 while even predicting some negative percentages. It's important to notice the lack of values in our fitted models after 0.7 as well, indicated that our model does an extremely poor job past that point in predicting the actual values >0.7.

1.4 Discussion

Overall betting odds are an accurate predictor of winning probability. However, for higher betting odds, they tend to overestimate the probability of winning. This may be due to a lower sample size among higher betting odds which can skew the data. This is valuable because this allows us to segway into question 2 with expectation in mind when examining other variables including these Odds. After finishing the exploratory analysis, we discovered that betting Odds and Ranking points

Odds. After finishing the exploratory analysis, we discovered that betting Odds and Ranking points were both valuable in predicting percent of matches won. Intuitively, it's clear as why they might be good predictors because as discussed in question 1, betting odds will mostly always favor that of those most likely to win and ranking points accumulates with through winning matches throughout a tournament so naturally both of these will be extremely correlated. Although a little more surprising, it's logical that home country wouldn't make too much of a difference upon win percentage because tennis courts could vary so drastically even in one's country from factors such as weather and court surface just to name a few. But ultimately, what we found most surprising was that height couldn't act as a valuable predictor. As we saw it, it seemed as though height advantage would produce various benefits such as vertical step speed due to stride lengths and arm length advantages. It's important to note that the R^2 of the model was particularily low meaning that despite having some statistically significant predictors, there was a significant amount of win percentages that simply couldn't be explained. This could be attributed to our original dataset containing nearly ~30 predictor variables whereas we only used 5 of those predictor variables. Additionally, sports are naturally very demanding and thus various factors from how one might be feeling that day, injuries, and so many unaccountable human factors that cannot be tracked nor documented.

When it comes to predicting win percentages of these matches, it's crutial to stress that they are all professional matches played among the very best tennis players around the globe. So naturally, these atheletes will most likely be more physical capable than the average person. So when compared among likewise atheletes, the range in something such as height may be more limited and may very well be above that of an regular person. Therefore, when measuring the proficiency of a tennis player, the height's of these atheletes might not be accurate in determining tennis performance overall. Thefore if we wanted to perform future research, in order to achieve more accurate predictions, we'd have to sample from a larger, more inclusive population that includes everyone rather than just professionals.